CORPUS

Corpus

7 | 2008
Constitution et exploitation des corpus d'ancien et de moyen français

Pierre Kunstmann, Achim Stein (éds.) — *Le Nouveau Corpus d'Amsterdam*

Bohdana Librova

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/1636
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 10 novembre 2008
ISSN : 1638-9808

Référence électronique
Bohdana Librova, « Pierre Kunstmann, Achim Stein (éds.) — *Le Nouveau Corpus d'Amsterdam* », *Corpus* [En ligne], 7 | 2008, mis en ligne le 13 novembre 2009, consulté le 04 mai 2019. URL : http://journals.openedition.org/corpus/1636

Ce document a été généré automatiquement le 4 mai 2019.

Pierre Kunstmann, Achim Stein (éds.) — *Le Nouveau Corpus d'Amsterdam*

Bohdana Librova

RÉFÉRENCE

Pierre Kunstmann, Achim Stein (éds.), *Le Nouveau Corpus d'Amsterdam, Actes de l'atelier de Lauterbad, 23-26 février 2006.* Stuttgart : Franz Steiner Verlag, 2007, Zeitschrift für Französische Sprache une Literatur, Neue Folge, Heft 34, 200 pages

1 Ce volume collectif réunit les communications présentées à l'atelier de Lauterbad, qui s'est tenu les 23-26 février 2006. Il fait état des derniers avancements dans le domaine des corpus électroniques du français médiéval, tout en renseignant sur les modes de leur exploitation.

2 Les quatre premières communications se penchent sur le Nouveau Corpus d'Amsterdam (désormais NCA).

3 Les paramètres fondamentaux du NCA sont exposés par Pierre Kunstmann et Achim Stein dans leur article liminaire : il s'agit d'un remaniement du corpus de textes littéraires en ancien français établi par l'équipe d'A. Dees au début des années 1980. A l'heure actuelle, le NCA réunit environ 300 textes et extraits de textes littéraires datés d'environ 1150-1350, lemmatisés et annotés grammaticalement.

4 Après avoir retracé l'histoire du premier corpus d'A. Dees, les auteurs exposent la méthodologie du NCA : l'étiquetage grammatical, partiellement effectué déjà par l'équipe de Dees, a été complété par une lemmatisation fondée sur plusieurs ressources, parmi lesquelles le Tobler-Lommatzsch joue un grand rôle. Les apories liées à la recherche des lemmes inspirent aux auteurs d'édifiantes considérations lexicologiques : c'est, en effet, par ce biais qu'ils ont pu constater des incohérences internes de sources lexicographiques, telles que l'absence, dans le Tobler-Lommatzsch, de distinctions

explicites entre les fonctionnements pronominal et adjectival des indéfinis, des démonstratifs et des possessifs (l'on peut néanmoins présumer que ce défaut tient plus aux spécificités de la terminologie grammaticale allemande qu'à une inadvertance des lexicographes).

5 Si les considérations sur la méthode quantitative et sur la lemmatisation réjouiront le lexicologue, le linguiste du corpus trouvera son bonheur dans l'exposé de l'étiquetage grammatical (compliqué par la nécessité de concilier les nouvelles étiquettes avec celles introduites par Dees). Il y lira en particulier le détail du fonctionnement de l'étiqueteur « TreeTagger » et les modalités de l'application au NCA du balisage XML.

6 Yves Charles Morin évoque les aspects dialectologiques du NCA. Rappelant les postulats dialectologiques d'A. Dees ainsi que sa méthode de localisation des textes littéraires à partir des chartes parfaitement localisées et datées, il informe des progrès obtenus sur ce champ par la nouvelle équipe. Il expose ensuite les problèmes liés à l'exploitation dialectologique du NCA : perte de fichiers informatiques, présence d'informations contradictoires sur la localisation de manuscrits, négligence des éditeurs face à l'indication des leçons rejetées, arbitraire de la résolution des abréviations, difficulté à mesurer la variation des traits dialectaux dans le temps faute d'une représentativité chronologique suffisante des chartes, enfin contamination linguistique de nombreux textes. Y. Ch. Morin attire l'attention sur la nécessité de la dissociation entre la graphie et la phonie lors de la localisation des documents : c'est uniquement la composante graphique – les usages graphiques étant notablement individualisés selon les régions – qui doit être retenuecomme critère.

7 La même spécialisation régionale des graphies qui facilite la localisation des manuscrits, dessert le linguiste soucieux de reconstruire les états de prononciation anciens, le même graphème pouvant noter des phonèmes différents dans des aires géographiques voisines. La reconstruction des prononciations anciennes peut être, en revanche, aidée par l'étude des articulations dialectales modernes, comme pour les prépositions contractées *as / aus*.

8 Il faut rectifier, selon l'auteur, le refus intransigeant opposé par Dees à la « scriptologie », hypothèse du rayonnement des systèmes scripturaux d'une région à l'autre. Le cas du digramme *ou* utilisé pour noter le [o]ouvert dans la région de Liège, de même que l'emploi massif du digramme *oi* dans des régions dans lesquelles la prononciation correspondante oscillait entre [ei] et [oi], voire se ramenait à [e(i)], semblent appuyer l'hypothèse. Aussi bien, ces considérations sur la complexité des liens entre la graphie et la phonie intègrent les données du NCA dans une démarche complexe sollicitant différentes approches et différents postulats théoriques en dialectologie ancienne et moderne.

9 Martin D. Glessgen et Xavier Gouvert font le point des lacunes et imperfections des données philologiques fournies par A. Dees, et proposent d'y porter remède. Il s'avère que, pour certaines œuvres, Dees n'avait pas accès à des éditions de qualité. Il lui est, en outre, arrivé de moderniser le texte des éditions diplomatiques, voire de commettre des erreurs de transcription. Des données fondamentales, telles que l'identité de l'auteur, la classification générique ou la datation de l'œuvre font fréquemment défaut dans le premier Corpus d'Amsterdam. La bibliographie est muette sur des paramètres philologiques importants, tels que les caractéristiques des manuscrits, le type des éditions utilisées et leur qualité. Au vu de ces inconséquences, Glessgen et Gouvert ont entrepris un travail visant à améliorer la fiabilité du Corpus. Ils ont proposé d'en éliminer les éditions les plus problématiques et d'en remplacer d'autres, obsolètes, par des éditions récentes. Ils projettent en outre de revoir les localisations proposées par Dees avec des

instruments nouveaux, tels que le corpus des *Plus anciens documents linguistiques de la France.* Bien qu'A. Dees ait assuré une bonne représentativité générique du corpus, de telle sorte que « les œuvres et les auteurs phares sont présents », certaines lacunes ont été constatées, et la part qui revient aux différents genres reste à rééquilibrer.

10 Les auteurs ont, en outre, éclairci la structure du NCA en introduisant quinze nouvelles données dans le fichier analytique, parmi lesquelles le lieu de la composition du texte et celui de la rédaction du manuscrit, la date du manuscrit et la qualité de l'édition.

11 On trouvera en annexe plusieurs exemples des fruits de leur travail, tels qu'une concordance entre les sigles de Dees et ceux, partiellement modifiés, du DEAF, un tableau indiquant le type (« diplomatique » ou « critique ») et la fiabilité des éditions utilisées par Dees, le type de saisie des textes (intégral, partiel, fragmentaire), la liste des éditions à retrancher ou à substituer, des passages à élargir et enfin un exemple d'une notice du NCA complétée.

12 Lene Schøsler préconise l'inclusion des manuscrits au NCA. La présence de ceux-ci permet de faire avancer les recherches dans trois domaines : 1) le suivi des modifications chronologiques et diatopiques (ces dernières étant déterminées à l'aide de l'Atlas des chartes d'A. Dees), 2) les recherches stemmatologiques, 3) le développement d'études diachroniques fondées sur la variation manuscrite.

13 L'intégration des manuscrits pose néanmoins un certain nombre de problèmes, tels que les modalités de transcription : faut-il introduire la ponctuation moderne, quel choix adopter face à la résolution des abréviations des mots ne se trouvant pas écrits en pleines lettres dans le manuscrit, faut-il corriger des erreurs manifestes, comment concilier, au sein de la base, la transcription diplomatique des manuscrits avec le texte des éditions, etc. L. Schøsler appuie son propos par deux exemples prouvant la pertinence de l'étude systématique des documentsmanuscrits : une étude comparée entre la variation des graphies repérées sous la plume d'un même scribe et celle constatée sous la plume de deux scribes contemporains, la conduit à conclure à l'absence d'une différence appréciable entre les deux cas de figure.

14 L'examen de la variation au sein d'une tradition manuscrite peut enfin révéler des régularités pour des phénomènes que l'on serait, au vue d'un seul témoin manuscrit, tenté de qualifier d'arbitraires, tel l'emploi des temps du passé. Ainsi, l'analyse des différentes versions manuscrites du *Charroi de Nîmes* a permis à l'auteur de dégager des valeurs propres à certains temps ainsi que leur distribution complémentaire.

15 La deuxième partie du volume (pp. 101-200) est dévolue à des bases de données autres que le Corpus d'Amsterdem, et à des cas d'application.

16 Hiltrud Gerner explique la structure et le fonctionnement des corpus textuels élaborés par l'ATILF, tout en insistant sur leur interconnexion avec d'autres ressources. Il prête une attention soutenue à leur histoire et aux choix méthodologiques ayant présidé à leur élaboration. L'article fournit des repères utiles sur l'accessibilité des corpus et sur les types de recherche qu'ils permettent d'effectuer. Il passe ainsi en revue le *Frantext moyen français*, sa continuation sous forme de *7FMR* et les trois bases du *Dictionnaire du moyen français,* avant d'expliquer comment les lemmes de ces dernières ont servi à G. Souvay pour l'élaboration du programme de lemmatisation *LGeRM* (*Lemme, Graphies et Règles Morphologiques*), précieux outil permettant d'identifier les formes graphiques occurrentes tout en épinglant les formes non encore traitées. Enfin, le moyen français sera intégré à une base à visée diachronique, la *Base de Connaissances Lexicales*, qui devrait réunir l'ensemble des ressources de l'ATILF sous un hyper-navigateur assurant des passages

entre les différents corpus. Il ne reste qu'à souhaiter un prochain achèvement de l'outil prometteur qui devrait assortir chaque « mot, forme, lemme » ou « étymon » de l'ensemble des informations morphologiques, lexicologiques et étymologiques « disponibles ».

17 Pierre Kunstmann et Gilles Souvay présentent leur *Dictionnaire électronique de Chrétien de Troyes* (le DECT, en accès libre en ligne sous sa forme provisoire, contenant les entrées des lexèmes commençant en A et B), en cours d'élaboration. C'est un outil rigoureux à toute épreuve : élaboré exclusivement à partir de manuscrits, lemmatisant et étiquetant l'intégralité de la matière du manuscrit de Guiot, sans pour autant négliger les concordances avec les variantes lexicales des autres codex, il permettra d'effectuer des allées et venues systématiques entre le dictionnaire et le corpus des textes de Chrétien. Toute forme graphique occurrente sera assortie d'un lemme et d'une annotation grammaticale. Une fois l'œuvre achevée, le lecteur pourra vérifier les graphies sur des images de manuscrits. Une attention particulière est portée à l'analyse de la combinatoire syntaxique (notamment à la valence verbale) et au classement onomasiologique : ce dernier facilite, entre autres, l'étude des séries synonymiques et des axes antonymiques.

18 Une seule considération critique m'a effleurée face à cet instrument qui semble accomplir les rêves les plus secrets de la lexicographie médiévale : malgré la difficulté liée à l'identification des degrés de figement dans un état de langue révolu, peut-être pourrait-on songer à un marquage lexicologique des locutions et des constructions à verbe support, qui ne semblent pour l'instant pas être distinguées des constructions de mots libres.

19 France Martineau, Constanta Rodica Diaconescu et Paul Hirschbühler informent du projet *Modéliser le changement : les voies du français,* orienté vers l'histoire du français canadien. Dans un premier temps, ce projet vise à l'élaboration d'un « corpus représentatif de la complexité des échanges dans la société française, du Moyen Âge au XVII^{e} siècle, pour le français d'Europe, [et] du début de la colonisation jusqu'aux années qui sont suivi la Conquête, pour la Nouvelle-France (XVII^{e} et XVIIIe siècles et début du XIX^{e} siècle) ». Dans un second temps, les auteurs espèrent pouvoir fonder sur ce corpus un nouveau modèle du changement linguistique, modèle qui tiendra compte des mécanismes de l'acquisition du langage et de l'histoire sociale. Une comparaison avec d'autres corpus de langues anciennes donne au projet une dimension de linguistique générale. La langue est enfin saisie dans sa fonction de l'enjeu identitaire, question liée à celle de la formation des normes régionales.

20 Quant à la structure du corpus, ont été privilégiés, pour des raisons liées à l'histoire linguistique du Canada, les dialectes du Nord-Ouest de la France. Les auteurs du projet s'attachent à y inclure des textes non littéraires, décision digne d'être saluée.

21 Les textes sont fidèlement transcrits à partir de leur support d'origine, qu'il s'agisse de manuscrits ou d'éditions sur papier. Le balisage du corpus traditionnel est enrichi de données concernant le scripteur, permettant d'associer des informations linguistiques à des faits d'ordre socio-historique.

22 Le procédé d'étiquetage morpho-syntaxique est expliqué à l'exemple de la composante médiévale du corpus : le lecteur peut jauger la pertinence du système d'annotation en considérant les traces d'éléments non exprimés ou déplacés, que le corpus permet de révéler. L'aptitude du corpus à nourrir des études en syntaxe historique ressort au vu d'une analyse de la non expression du pronom personnel sujet.

23 De l'aveu des auteurs, la conception de la base serait facilitée par le recours à des auxiliaires tels qu'un « lexique lemmatisé des formes de l'époque ». En ce sens, ils appellent à une collaboration entre équipes de recherche et suggèrent de concevoir un système d'exploitation croisée de différents corpus, notamment des corpus textuels et des bases de données lemmatisées telles que le NCA.

24 Céline Guillot, Alexei Lavrentiev et Christiane Marchello-Nizia exposent les structures, fonctions et perspectives de la *Base de Français Médiéval* (BFM). Vu la priorité accordée aux recherches en syntaxe, domaine requérant une masse de données considérable, la décision a été prise d'exploiter des éditions fiables, bien que la transcription « quasi-diplomatique » des manuscrits ait été également entamée, avec pour objectif de fournir un support à l'évaluation de la qualité des éditions critiques, et de documenter des études des graphies et de la ponctuation.

25 On sait l'attention soutenue portée par les concepteurs de la Base aux « métadonnées » (les annotations morpho-syntaxiques et, dans un proche avenir, également sémantiques), inappréciables pour l'analyse linguistique des textes, mais nombreuses et variées au point que les néophytes hésitent parfois à utiliser l'ensemble des fonctionnalités de l'interface WEBLEX, se limitant à des requêtes lexicales basiques et profitant de l'amabilité de l'équipe pour commander des recherches selon des critères plus poussés. Or, si le lecteur - comme c'est le cas de l'auteur des présentes lignes - se sentait jusqu'ici quelque peu désemparé face aux multiples fonctionnalités de WEBLEX, le présent article lui fournit quelques certitudes : une version succincte des grands principes qui pourront le guider lors de sa recherche dans les bases, un résumé de données concernant les spécificités des différentes versions – des BFM 1, 2 et 3, et même de la BFM 0 – ainsi que quelques suggestions de types de recherche qu'il pourra effectuer grâce à l'astucieux balisage XML - TEI. L'utilisateur de la base est en même temps mis en garde contre des imperfections – sans doute temporaires – de certaines versions.

26 Quant aux perspectives ébauchées, l'on peut se réjouir de la volonté d'inclure à la base des textes non littéraires, et de rééquilibrer la part qui y revient aux différents genres et époques (à ce propos, on pourrait suggérer d'intégrer quelques textes de nature plus austère, à visée moralisatrice et didactique, genre peu représenté dans les corpus, et d'en profiter pour corriger leurs éditions vétustes). En écho à l'article précédent, la conclusion ouvre sur une visée de mutualisation des démarches métalinguistiques : il serait en effet souhaitable que les différentes équipes puissent convenir d'un noyau commun, qui serait exploitable selon des critères de recherche identiques, quelle que soit la base interrogée.

27 David Trotter présente la base de données des textes anglo-normands ainsi que l'Anglo-Normand Dictionary (déjà paru entre 1977 et 1992), en cours de révision (désormais AND). Le corpus anglo-normand a été conçu selon des critères philologiques exemplaires : ont été préférés des textes soit inaccessibles, soit intéressants au point de vue lexical. C'est ainsi que le corpus offre la version numérisée intégrale des textes publiés sous l'égide de l'Anglo-Norman Text Society, lexicologiquement intéressants et relativement mal accessibles.

28 Le corpus sert de base à l'AND, tout en élargissant sa portée : l'étroite interconnexion entre le dictionnaire et le corpus permet de circuler aisément entre les citations de l'AND et le texte du corpus : en un seul double-click, on passe d'une occurrence d'un terme située dans un texte (figurant dans le corpus ou même simplement sur la toile) à l'entrée

de l'AND, ainsi qu'à toutes les occurrences du terme présentes dans le dictionnaire. Le corpus permet également de mettre en évidence des collocations de séquences figées.

29 L'auteur enfin s'associe aux voix qui s'étaient fait entendre à plusieurs reprises dans ce volume, en prônant une unification des systèmes d'annotation des différents corpus : en effet, une telle décision faciliterait la comparaison entre l'anglo-normand et la français médiéval commun, une des principales ambitions du projet.

30 Martin Elsig et Esther Rinke mettent le NCA à l'épreuve d'une investigation ponctuelle : il s'agit d'étudier l'inversion du sujet dans les propositions dont la première zone est occupée par des adverbes. Après avoir défini leur corpus selon des critères philologiques pertinents, les auteurs constatent que le balisage du NCA permet de répondre aux exigences philologiques les plus rigoureuses.

31 Ils expliquent ensuite la démarche adoptée lors de l'interrogation du corpus, et présentent les résultats classés automatiquement en fonction de différents critères présélectionnés (tels que répartition dialectale et générique des données, statistique des adverbes placés en position frontale etc.). En dépit de la nécessité de quelques manipulations manuelles (exclusion des propositions à sujet nul), la base s'avère hautement opérante.

32 Les résultats statistiques demandent néanmoins à être affinés par une considération linguistique, qu'aucun logiciel ne saurait opérer : ainsi les auteurs ont-il cru observer une étroite parenté entre les fonctions du sujet antéposé au verbe, et celles du sujet nul, dans la dynamique informationnelle de l'énoncé. Selon eux, en effet, dans les deux cas, le sujet assumerait la fonction de topique, par opposition au sujet postposé, qui lui serait rhématique : ce résultat les amène à déclarer que le sujet nul occupe une place préverbale virtuelle – typique des éléments thématiques – et que, par conséquent, les adverbes ne déclanchent la postposition du sujet que dans environ 50% des cas, ce qui, à leurs yeux, conduit à relativiser le statut de l'ancien français en tant que langue à V2.

33 L'ouvrage s'achève sur trois interventions qui avaient été présentées dans une table ronde consacrée au corpus des chartes.

34 Yves Charles Morin se penche sur la morphologie du verbe dans les chartes. Il met en garde contre l'utilisation des chartes comme source unique en linguistique (pour la morphologie du verbe, en particulier, la matière lexicale y est trop peu diversifiée ; on y observe, de surcroît, l'influence des traditions scripturales extérieures à la région donnée). Par conséquent, en vue d'offrir des résultats linguistiquement fiables, l'étude des chartes doit être associée à d'autres outils de la dialectologie historique.

35 Hans Goebl, quant à lui, annonce avoir découvert, dans l'*Atlas des chartes* de Dees, certaines structures profondes, sans doute insoupçonnées par Dees lui-même. La comparaison quantitative entre celles-ci et l'ALF lui a permis de constater une remarquable continuïté des « assises diatopiques » de la région d'Oïl entre le 13^{e} siècle et la période moderne.

36 L'auteur prend ensuite position vis-à-vis du problème de la correspondance entre la graphie et la phonie, en proposant de voir dans chaque texte médiéval la résultante de plusieurs forces « antagonistes » : des traditions graphiques d'une part, et, d'autre part, de la prononciation effective propre à chaque région : c'est le « substrat dialectal générateur », déjà connu des œuvres de l'auteur.

37 HG fait enfin une plaidoirie vigoureuse en faveur des approches quantitatives dans la linguistique du corpus, et rappelle la nécessité d'appliquer des méthodes mathématiques

rigoureuses pour la détermination du taux de dialectalité des textes médiévaux, par définition graphiquement « hybrides ».

38 Enfin, Yuji Kawaguchi propose une réflexion sur l'état des méthodes en dialectologie médiévale, en appuyant son propos par une analyse des actes écrits en champenois au 13e siècle. Il prévient contre d'apparentes certitudes représentées par les coordonnées officielles des chartes (les chartes sont souvent rédigés par des scribes provenant de régions différentes de celles où elles sont établies, il faut se méfier des cartulaires etc.), et relativise ainsi leur exploitabilité dialectologique. Il met en évidence la complexité de la scripta médiévale telle qu'elle se manifeste dans les chartes, tiraillée, selon lui, entre une tendance à l'uniformisation des graphies selon un étalon « suprarégional », et la nécessité de noter les formes dialectales propres à la région de la rédaction du document.

39 Cette hétérogénéité des graphies n'a pas empêché Y. Kawaguchi d'aboutir à des résultats impressionnants concernant l'augmentation graduelle de la graphie « ou » pour noter le produit du changement phonétique [o] > [u] (on souhaiterait néanmoins disposer de plus de détails sur l'environnement phonétique de la voyelle concernée). L'auteur termine en insistant sur l'importance de la dimension diachronique en dialectologie.

40 Il reste à conclure. Nous en profiterons pour rappeler en trois points les grands apports de l'ouvrage (on nous pardonnera d'en passer d'autres sous silence). Outre l'intérêt pratique que présentent les informations sur le contenu textuel des bases, leur accessibilité et leurs diverses fonctionnalités, le volume a le mérite d'exposer en détail les choix philologiques et méthodologiques qui ont guidé leur réalisation. Cette perspective métalexicographique rend l'ouvrage doublement utile : d'une part, il sensibilise le linguiste à l'ensemble des critères dont il faut idéalement tenir compte lors de la conception des bases de données et des dictionnaires consacrés à la langue médiévale; d'autre part, il avertit l'utilisateur des lacunes et approximations, provisoires pour certaines bases, inévitables pour d'autres. En mettant ainsi à nu l'architecture des corpus, l'ouvrage invite les grammairiens à des recherches de qualité.